

Learning with Cross-Kernels and Ideal PCA

Franz J. Király*

Martin Kreuzer[†]

Louis Theran[‡]

Abstract

We describe how cross-kernel matrices, that is, kernel matrices between the data and a custom chosen set of ‘feature spanning points’ can be used for learning. The main potential of cross-kernels lies in the fact that (a) only one side of the matrix scales with the number of data points, and (b) cross-kernels, as opposed to the usual kernel matrices, can be used to certify for the data manifold. Our theoretical framework, which is based on a duality involving the feature space and vanishing ideals, indicates that cross-kernels have the potential to be used for any kind of kernel learning. We present a novel algorithm, Ideal PCA (IPCA), which cross-kernelizes PCA. We demonstrate on real and synthetic data that IPCA allows to (a) obtain PCA-like features faster and (b) to extract novel and empirically validated features certifying for the data manifold.

1. Introduction

Since their invention by Boser, Guyon and Vapnik [2, 7], kernel methods have had a fundamental impact on the fields of statistics and machine learning. The major appeal of using kernel methods for learning consists in using the kernel trick, first proposed by Aizerman, Braverman and Rozonoer [1], which allows to make otherwise costly computations in the feature space implicit and thus highly efficient for a huge variety of learning tasks – see e.g. [4, 6] for an overview.

Many kernel methods make extensive use of the so-called kernel matrix, a matrix whose entries are kernel function evaluations $k(x_i, x_j)$ at data points $x_1, \dots, x_N \in \mathbb{R}^n$. This kernel matrix $K_{XX} = (k(x_i, x_j))_{ij}$ is simultaneously the main source of efficient linearization and the central computational bottleneck. For instance, the most expensive part of algorithms such as kernel PCA or kernel ridge regression consists in computing the inverse or a singular value decomposition of the $(N \times N)$ matrix K_{XX} . We make two observations in this context:

(A) Subsampling K_{XX} , i.e., considering $(N \times M)$ sub-matrices of K_{XX} , is the state-of-the-art in speeding up the scalability in the number of data points N . While much has been written on the topic, there seems to be no consensus on how exactly to choose the subsample from the data set.

(B) There seems to be no kernel algorithm which learns the data manifold from which the x_i were obtained, or test whether an unseen data point is on the manifold - in particular none involving K_{XX} . (Note that kernel PCA outputs projections on predominant data coordinates, but *not* the coordinates embedded in data space or the shape of the data manifold.)

*Department of Statistical Science, University College, Gower Street, London WC1E 6BT, United Kingdom, f.kiraly@ucl.ac.uk

[†]Faculty of Computer Science and Mathematics, University of Passau, 94030 Passau, Germany, martin.kreuzer@uni-passau.de

[‡]Institut für Mathematik, Freie Universität Berlin, Arnimallee 2, 14195 Berlin, Germany, theran@math.fu-berlin.de

In this paper, we propose an algebraic duality framework which offers both a potential explanation and a potential solution to the above (seemingly unrelated) issues. Our contribution is centered around *cross-kernel matrices* $K_{XZ} = (k(x_i, z_j))_{ij}$, where $z_1, \dots, z_M \in \mathbb{R}^n$, called *cross-kernel basis points*, are not necessarily data points. Assuming that z_1, \dots, z_M are chosen suitably (e.g., randomly) and that M is large enough, we show theoretically and empirically that:

(1) The cross-kernel matrix K_{XZ} can be used to capture both the information contained in K_{XX} as well as additional information about the data manifold which is provably not contained in K_{XX} . Practically, we present an algorithm, called IPKA, which is able to learn both *coordinates in the manifold* as well as *certificates for being contained in the data manifold*, addressing issue (B) above.

(2) The cross-kernel matrix K_{XZ} can be employed in ways completely analogous to a subsampled matrix K_{XX} . As opposed to subsampling, the properties of the points z_i can be prescribed by the experimenter and are thus independent of known or unknown – and potentially detrimental – properties of the data x_i . In particular, K_{XZ} can be used for speeding up kernel learning, while potentially avoiding data-related subsampling issues, thereby addressing issue (A) above.

In the following we briefly explain why and how considering the cross-kernel K_{XZ} can be advantageous as compared to the kernel matrix K_{XX} . More detailed technical statements and explicit algorithms can be found in section 2 and thereafter.

Why the kernel matrix is not enough. The kernel matrix K_{XX} misses information on the data manifold that can be obtained from K_{XZ} . Let us explain this in the example where the kernel is the ordinary Euclidean scalar product $k(x, y) = \langle x, y \rangle$ in \mathbb{R}^n . Suppose the data points x_1, \dots, x_N all lie in a vector subspace $L \subseteq \mathbb{R}^n$. The entries of the kernel matrix K_{XX} are scalar products of the type $\langle x_i, x_j \rangle$ which are invariant under rotation of the coordinate system. Therefore, one obtains one and the same kernel matrix K_{XX} when considering rotated data points Ux_1, \dots, Ux_N and the rotated vector space $U \cdot L$ for any rotation matrix $U \in \mathbb{R}^{n \times n}$. Since one and the same kernel matrix K_{XX} can arise in a non-degenerate way from every vector space L (of that dimension), the data manifold L can not be obtained back from K_{XX} . Similar considerations hold for arbitrary kernels: *the data manifold can not be learnt from the kernel matrix K_{XX} .*

Cross-kernels help! In the above example the vector space L can be identified from a cross-kernel matrix $K_{XZ} = (k(x_i, z_j))_{ij}$, if z_1, \dots, z_M span \mathbb{R}^n . Observe that the rows of K_{XZ} are coordinate representations of the x_i . Since z_1, \dots, z_M span \mathbb{R}^n , the data point x_i can be obtained back from the i -th row of K_{XZ} . Since reconstructing x_i from the i -th row of K_{XZ} is a linear reparameterization, an unseen data point x is contained in the data manifold L if and only if the vector $\kappa_z(x) = (k(x, z_j))_j$ is contained in the row-span of K_{XZ} . Therefore we can obtain L back from K_{XZ} and use the row-span of K_{XZ} to efficiently test membership for the data manifold L . Note that the same reasoning does not work for the matrix K_{XX} , since x_1, \dots, x_N do not span \mathbb{R}^n , but only L . A similar reasoning holds in larger generality: *the cross-kernel K_{XZ} contains extensive information on the data manifold which K_{XX} does not.*

From cross-kernels to kernel learning. The “manifold awareness” of the cross-kernel implies further interesting properties which can be used for decomposing, and thus approximating, the original kernel matrix K_{XX} via the potentially much smaller K_{XZ} . Namely, letting $K_{ZX} = K_{XZ}^\top$ and $K_{ZZ} = (k(z_i, z_j))_{ij}$, we can show that the equality $K_{XX} = K_{XZ} K_{ZZ}^{-1} K_{ZX}$ holds *exactly* for polynomial kernels (assuming “feature-spanning” points z_i as in the example above, see Theorem 3.4). We conjecture (and have observed empirically) that the same holds approximately for other kernels. This equality is reminiscent of the central equation in subsampling, with the difference that

the points z_j do not need to be chosen among the data. Suitable choices of the points z_j avoid potential problems in the data while leading to the same speed-up of arbitrary kernel methods.

2. Some dualities for polynomial kernels

The following preview of our approach is a more technically exact and quantitative variant of the discussion in the introduction. In what follows, we use inhomogeneous polynomial kernels. (A further difficulty, not considered here, lies in the generalization to other kernels and kernel feature spaces.) The central observation which we carry over from the introduction is the certifying property of the row-span of K_{XZ} . In our setup, the data points x_i are sampled from a manifold $\mathcal{X} \subset \mathbb{R}^n$ which is “cut out” by an unknown set of polynomial equations. Our goal is to relate these equations to the linear span of the images $\Phi_{\leq d}(x_i)$ under the feature map $\Phi_{\leq d}$. This linear span is then related to the row-span of K_{XZ} by applying several dualities in and on feature space.

More technically, we first show how to identify $\mathbb{R}[\mathbf{t}]_{\leq d}$, the set of polynomials of degree at most d , with linear functionals on the kernel feature space and that every $f \in \mathbb{R}[\mathbf{t}]_{\leq d}$ can be expressed as a kernel decision function. To understand the equations cutting out \mathcal{X} (which are not canonical) it is better to change perspective and consider the set of *all* polynomials vanishing on \mathcal{X} , called its *vanishing ideal* (which is canonical). The main results of this section say, informally, that \mathcal{X} can be identified with the intersection of a particular linear space L and $\Phi_{\leq d}(\mathbb{R}^n)$, and that the vanishing ideal of \mathcal{X} corresponds to L^\perp , which in turn is isomorphic to the row-span of K_{XZ} . In other words, we relate a linear duality to the algebraic-geometric concept of duality between sets of equations and their common vanishing locus.

2.1. Duality of polynomial rings and feature spaces

To begin with, we introduce inhomogeneous polynomial kernel as follows.

Definition 2.1. Let $\theta \in (0, 1)$ be a fixed real number, and let $d \geq 1$. The *inhomogeneous polynomial kernel function* $k_{\leq d} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ is given by $k_{\leq d}(x, y) = (\theta \cdot \langle x, y \rangle + 1)^d$, where $\langle \cdot, \cdot \rangle$ denotes the standard scalar product.

This definition differs slightly from the usual one which is obtained after dividing by θ^d . Since θ is chosen arbitrarily in $(0, 1)$, no qualitative change is introduced by our convention. It is however, as we will see, the more natural one.

It is well-known that *feature space* $\mathcal{F}_{\leq d}$ of $k_{\leq d}$ satisfies $\mathcal{F}_{\leq d} \cong \mathbb{R}^m$, where $m = \binom{n+d}{d}$, and that the *feature map* $\Phi_{\leq d} : \mathbb{R}^n \rightarrow \mathcal{F}_{\leq d}$ is given by $\Phi_{\leq d}(x) = (\gamma_\alpha x^\alpha \mid |\alpha| \leq d)$ (e.g., see [4], Sec. 2.1). Here we let $x^\alpha = c_1^{\alpha_1} \cdots c_n^{\alpha_n}$ and $|\alpha| = \alpha_1 + \cdots + \alpha_n$ for $x = (c_1, \dots, c_n)$ and $\alpha = (\alpha_1, \dots, \alpha_n)$. Moreover, we have $\gamma_\alpha = \sqrt{\theta^{|\alpha|} \cdot \binom{d}{|\alpha|} \cdot \binom{|\alpha|}{\alpha_1, \dots, \alpha_n}}$. To define $\Phi_{\leq d}$ uniquely, we order p_1, \dots, p_m increasingly with respect to the degree-lexicographic term ordering. The feature map is characterized by the property that $k_{\leq d}(x, y) = \langle \Phi_{\leq d}(x), \Phi_{\leq d}(y) \rangle$ for $x, y \in \mathbb{R}^d$, i.e., by the fact that it transforms the kernel function to a standard scalar product on the feature space.

On the algebraic side, the main objects linked via duality are vector spaces of polynomials. Let $\mathbf{t} = (t_1, \dots, t_n)$ be a tuple of indeterminates, and let $\mathbb{R}[\mathbf{t}]$ be the polynomial ring in these indeterminates. For every $d \geq 0$, we denote by $\mathbb{R}[\mathbf{t}]_{\leq d}$ the vector space of all polynomials of

degree at most d . Recall that the dimension of $\mathbb{R}[\mathbf{t}]_{\leq d}$ is $\binom{n+d}{d}$. The inhomogeneous polynomial kernel function can be extended to a pairing of polynomials as follows.

Notation 2.2. Let $\theta \in (0, 1)$ and $d \geq 1$. The inhomogeneous polynomial kernel on $\mathbb{R}[\mathbf{t}]$ is the map $\mathbf{k}_{\leq d} : \mathbb{R}[\mathbf{t}]^n \times \mathbb{R}[\mathbf{t}]^n \rightarrow \mathbb{R}[\mathbf{t}]$ given by $\mathbf{k}_{\leq d}(F, G) = (\theta \cdot \langle F, G \rangle + 1)^d$ where we define $\langle F, G \rangle = f_1 g_1 + \dots + f_n g_n$ for $F = (f_1, \dots, f_n)$ and $G = (g_1, \dots, g_n)$.

The polynomials $\mathbf{k}_{\leq d}(x, \mathbf{t}) \in \mathbb{R}[\mathbf{t}]$ can be used to generate $\mathbb{R}[\mathbf{t}]_{\leq d}$, as our next proposition shows. By the multinomial theorem, they can be expressed as $\mathbf{k}_{\leq d}(x, \mathbf{t}) = (\theta \langle x, \mathbf{t} \rangle + 1)^d = \sum_{|\alpha| \leq d} \gamma_\alpha^2 x^\alpha t^\alpha$. Recall that points $x_1, \dots, x_m \in \mathbb{R}^n$ are called *generic* for a property \mathcal{P} if there exists a Zariski open subset¹ of $U \subseteq (\mathbb{R}^n)^m$ such that property \mathcal{P} holds for all $(x_1, \dots, x_m) \in U$.

Proposition 2.3. Let $d \geq 0$, let $m' \geq \binom{n+d}{d}$, and let $x_1, \dots, x_{m'} \in \mathbb{R}^n$ be generic. Then we have $\mathbb{R}[\mathbf{t}]_{\leq d} = \langle \mathbf{k}_{\leq d}(x_i, \mathbf{t}) \mid 1 \leq i \leq m' \rangle$.

The next proposition says that $\mathbb{R}[\mathbf{t}]_{\leq d}$ is dual to the feature space $\mathcal{F}_{\leq d}$.

Proposition 2.4. (a) For $d \geq 1$, the map $\varphi : \mathbb{R}[\mathbf{t}]_{\leq d} \rightarrow (\mathcal{F}_{\leq d})^\vee \cong \text{Hom}_{\mathbb{R}}(\mathbb{R}^m, \mathbb{R})$ defined by $f = \sum_{|\alpha| \leq d} c_\alpha \mathbf{t}^\alpha \mapsto f^\vee$, where $f^\vee(e_\alpha) = c_\alpha / \gamma_\alpha$ is an isomorphism of \mathbb{R} -vector spaces. (b) For $d \geq 1$, the map $\psi : \mathbb{R}[\mathbf{t}]_{\leq d} \rightarrow \mathcal{F}_{\leq d}$ defined by $f = \sum_{|\alpha| \leq d} c_\alpha \mathbf{t}^\alpha \mapsto \sum_{|\alpha| \leq d} (c_\alpha / \gamma_\alpha) e_\alpha$ is an isomorphism of \mathbb{R} -vector spaces.

By dualizing the map φ and using the canonical isomorphism between $\mathcal{F}_{\leq d}$ and its bidual, we obtain an isomorphism $\varphi^\vee : \mathcal{F}_{\leq d} \rightarrow (\mathbb{R}[\mathbf{t}]_{\leq d})^\vee$ which maps e_α to the \mathbb{R} -linear map given by $t^\beta \mapsto \gamma_\alpha \delta_{\alpha\beta}$. Another consequence of the preceding proposition is that, if we pass to the union $\mathbb{R}[\mathbf{t}] = \bigcup_{d \geq 1} \mathbb{R}[\mathbf{t}]_{\leq d}$, we see that the polynomial ring contains the duals of all feature spaces $\mathcal{F}_{\leq d}$.

Finally, we interpret the isomorphism φ in terms of kernel decision functions. Recall that the map $\text{eval}(f) : \mathbb{R}^n \rightarrow \mathbb{R}$ given by $x \mapsto f(x)$ is called the polynomial function associated to $f \in \mathbb{R}[\mathbf{t}]$ and that the polynomial function associated to $\mathbf{k}_{\leq d}(x, \mathbf{t})$ is called a *kernel decision function*. Let us denote the vector space of all kernel decision functions by $\text{KDF}_{\leq d}$.

Corollary 2.5. The map $\Phi_{\leq d}^* : (\mathcal{F}_{\leq d})^\vee \rightarrow \text{KDF}_{\leq d}$ given by $\ell \mapsto \ell \circ \Phi_{\leq d}$ is an isomorphism.

Now we use the isomorphism $\varphi : \mathbb{R}[\mathbf{t}]_{\leq d} \rightarrow (\mathcal{F}_{\leq d})^\vee$ to transfer the standard scalar product on $\mathcal{F}_{\leq d}$ to $\mathbb{R}[\mathbf{t}]_{\leq d}$ in the natural way. The result is a scalar product $\langle \cdot, \cdot \rangle_\varphi$ on $\mathbb{R}[\mathbf{t}]_{\leq d}$ such that

$$\langle \mathbf{t}^\alpha, \mathbf{t}^\beta \rangle_\varphi = \langle (1/\gamma_\alpha) e_\alpha^\vee, (1/\gamma_\beta) e_\beta^\vee \rangle = (1/\gamma_\alpha^2) \delta_{\alpha\beta}$$

for $|\alpha| \leq d$. The next proposition provides a basic property of the scalar product $\langle \cdot, \cdot \rangle_\varphi$.

Proposition 2.6. For $f \in \mathbb{R}[\mathbf{t}]_{\leq d}$ and $x \in \mathbb{R}^n$, we have $f(x) = \langle f, \mathbf{k}_{\leq d}(x, \mathbf{t}) \rangle_\varphi$.

The duality expressed by the map φ is an algebraic analogue of the theory of reproducing kernel Hilbert spaces. The associated Hilbert space is the space of polynomial functions $f : \mathbb{R}^n \rightarrow \mathbb{R}$ of degree $\leq d$ which can be identified with $\mathbb{R}[\mathbf{t}]_{\leq d}$ by replacing the polynomial function f with the corresponding symbolic polynomial. The equation in the preceding proposition could also be obtained by combining the Riesz representation with this identification (see [4], Sec. 2.2). In the next section we go beyond what can be shown using the usual RKHS duality alone.

¹ All such sets have full measure under any continuous probability density.

2.2. Duality of vanishing ideals and feature spans

Next we show that ideals – a classical concept in algebra – are the proper dual objects of feature spans of manifolds, in the same way as the polynomial ring is the dual of feature space itself. Recall that an *ideal* I in $\mathbb{R}[\mathbf{t}]$ is a vector subspace such that $I \cdot \mathbb{R}[\mathbf{t}] \subseteq I$. Ideals are connected to subsets of \mathbb{R}^n as follows.

Definition 2.7. (a) Given a subset \mathcal{X} of \mathbb{R}^n , the set of polynomials $I(\mathcal{X}) = \{f \in \mathbb{R}[\mathbf{t}] \mid f(x) = 0 \text{ for all } x \in \mathcal{X}\}$ is an ideal in $\mathbb{R}[\mathbf{t}]$. It is called the *vanishing ideal* of \mathcal{X} .

(b) Given an ideal I in $\mathbb{R}[\mathbf{t}]$, the set of points $\mathcal{V}(I) = \{x \in \mathbb{R}^n \mid f(x) = 0 \text{ for all } f \in I\}$ is called the *zero set* of I . A subset \mathcal{X} of \mathbb{R}^n is called an *algebraic set* if it is the zero set of an ideal in $\mathbb{R}[\mathbf{t}]$.

Notice that not every ideal in $\mathbb{R}[\mathbf{t}]$ is a vanishing ideal, since vanishing ideals have the additional property of being *radical*, i.e., if $f \in I^i$ for some $i \geq 1$ then $f \in I$. Given an ideal I in $\mathbb{R}[\mathbf{t}]$ and $d \geq 0$, we let $I_{\leq d} = I \cap \mathbb{R}[\mathbf{t}]_{\leq d}$.

From now on the data manifold \mathcal{X} is always assumed to be an algebraic set in \mathbb{R}^n . In this case, the vanishing ideal $I(\mathcal{X})$ is dual to the manifold \mathcal{X} in the following sense.

Theorem 2.8. Let $\mathcal{X} \subset \mathbb{R}^n$ be an algebraic set.

- (a) We have $I(\mathcal{X})_{\leq d} = \langle \mathbf{k}_{\leq d}(x, \mathbf{t}) \mid x \in \mathcal{X} \rangle^\perp$, where \perp is taken with respect to $\langle \cdot, \cdot \rangle_\varphi$.
- (b) Under the isomorphism φ , the set $I(\mathcal{X})_{\leq d}$ corresponds to $\{\ell \in (\mathcal{F}_{\leq d})^\vee \mid \ell(\langle \Phi_{\leq d}(\mathcal{X}) \rangle) = 0\}$.
- (c) Under the isomorphism ψ , the set $I(\mathcal{X})_{\leq d}$ corresponds to the feature span $\langle \Phi_{\leq d}(\mathcal{X}) \rangle$.

The space $\langle \Phi_{\leq d}(\mathcal{X}) \rangle$ will be called the *feature span* of \mathcal{X} . As discussed above, Theorem 2.8 is a kernelized version of the usual algebra-geometry duality based on Hilbert’s Nullstellensatz.

Remark 2.9. All the results in this section hold, *mutatis mutandis*, for homogeneous polynomial kernels. For other kernels, they can be adapted via kernelizing the polynomial ring, but then exact statements may need to be replaced with approximate ones. For this reason, and to save space, we work with inhomogeneous polynomial kernels from now on.

3. Vanishing ideals and cross-kernel matrices

In this section we discuss structural and algebraic properties of cross-kernel matrices between data points and further sampled points. As in the previous section we use inhomogeneous polynomial kernels. Similar statements can be proven for homogeneous polynomial kernels, and even non-polynomial kernels. In the latter case our exact statements have to be transformed into spectral approximation results.

The following setting is used throughout the section. There is a manifold $\mathcal{X} \subseteq \mathbb{R}^n$ from which data points are sampled, and a manifold $\mathcal{Z} \subseteq \mathbb{R}^n$ from which further points are sampled in such a way that they feature-span \mathcal{Z} . We assume that both \mathcal{X} and \mathcal{Z} are algebraic sets and that there is a degree $d \geq 1$ such that \mathcal{X} and \mathcal{Z} are cut out by polynomials of degree at most d . Furthermore, we assume $\mathcal{X} \subseteq \mathcal{Z}$. The data manifold \mathcal{X} is considered to be fixed and unknown, while \mathcal{Z} can be chosen by the experimenter and is typically given by $\mathcal{Z} = \mathbb{R}^n$. The sampled data points are denoted by $x_1, \dots, x_n \in \mathcal{X}$, and the points samples from \mathcal{Z} are denoted by $z_1, \dots, z_M \in \mathcal{Z}$. They are the rows of the matrices $X \in \mathbb{R}^{n \times n}$ and $Z \in \mathbb{R}^{M \times n}$, respectively. We begin by introducing kernel and cross-kernel matrices.

Definition 3.1. (a) Given $d \geq 1$ and the inhomogeneous polynomial kernel $k_{\leq d} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$, we denote the matrix of size $N \times M$ whose entry in position (i, j) is $k_{\leq d}(x_i, z_j)$ by $K_{\leq d}(X, Z)$. This matrix is called the *cross-kernel matrix* between X and Z .

(b) The matrix $K_{\leq d}(X, X)$ is simply called the *kernel matrix* of X .

In the following we study properties of the matrices $K_{\leq d}(X, X)$, $K_{\leq d}(X, Z)$ and $K_{\leq d}(Z, Z)$, and we prove duality statements between cross-kernel matrices and vanishing ideals. An important condition is *genericity* of the points sampled from \mathcal{Z} . In our case, the appropriate genericity condition is that the points feature-span \mathcal{Z} . It is defined as follows.

Definition 3.2. Let $d \geq 1$, let $\Phi_{\leq d} : \mathbb{R}^n \rightarrow \mathcal{F}_{\leq d}$ be the feature map, and let $z_1, \dots, z_M \in \mathcal{Z}$.

(a) For a set $S \subseteq \mathbb{R}^n$, the \mathbb{R} -linear span $\text{fspan}(S) = \langle \Phi_{\leq d}(s) \mid s \in S \rangle$ is called the *feature span* of S , and the number $\text{frk} S = \dim_{\mathbb{R}}(\text{fspan}(S))$ is called the *feature rank* of S .

(b) We say that $z_1, \dots, z_M \in \mathcal{Z}$ *feature-span* the algebraic set \mathcal{Z} if $\text{fspan}(\{z_1, \dots, z_M\}) = \text{fspan}(\mathcal{Z})$. By a slight abuse of notation, we also say that the matrix Z feature-spans \mathcal{Z} in this case, and we write $\text{frk}(Z)$ for $\text{frk}(\{z_1, \dots, z_M\})$. Moreover, we say that z_1, \dots, z_M are *feature independent* if $\Phi_{\leq d}(z_1), \dots, \Phi_{\leq d}(z_M)$ are linearly independent in $\mathcal{F}_{\leq d}$.

The matrix Z is also called the *feature generating matrix*. Notice that, by elementary linear algebra, a set of points $\{z_1, \dots, z_M\}$ feature-spans \mathcal{Z} if and only if $\text{frk}(Z) = \text{frk}(\mathcal{Z})$. A traditional method to analyse the kernel matrix $K_{\leq d}(X, X)$ is to *subsample* X , i.e., to choose the points z_i in $\{x_1, \dots, x_N\}$. This entails the common problem that one cannot assume that Z feature-spans \mathcal{X} . Our next theorem characterizes the matrices Z which avoid this problem.

Theorem 3.3. *In the above setting, assume that $\text{fspan}(X) \subseteq \text{fspan}(Z)$.*

(a) (Cross-Kernel Rank) *We have $\text{rank} K_{\leq d}(X, Z) = \text{frk}(X)$. Hence, if X feature-spans \mathcal{X} then $\text{rank} K_{\leq d}(X, Z) = \text{frk}(\mathcal{X})$.*

(b) (Cross-Kernel Data Manifold Certificate) *Let $c \in \mathbb{R}^n$ such that $\Phi_{\leq d}(c) \in \text{fspan}(Z)$. Then we have $\Phi_{\leq d}(c) \in \text{fspan}(X)$ if and only if $(k_{\leq d}(c, z_1), \dots, k_{\leq d}(c, z_M))$ is contained in the row span of $K_{\leq d}(X, Z)$. If X feature-spans \mathcal{X} then this implies $c \in \mathcal{X}$.*

(c) (Cross-Kernel Nullspace) *Let $f \in I(\mathcal{Z})_{\leq d}^\perp$ and suppose that X feature-spans \mathcal{X} and Z feature-spans \mathcal{Z} . Then we have $f \in I(\mathcal{X})_{\leq d}$ if and only if f is of the form $f = \sum_{i=1}^M c_i k_{\leq d}(z_i, \mathbf{t})$ with a coefficient tuple $c = (c_1, \dots, c_M) \in \mathbb{R}^M$ such that $K_{\leq d}(X, Z) \cdot c^\top = 0$.*

(d) (Cross-Kernel Range) *Let $f \in I(\mathcal{Z})_{\leq d}^\perp$. Suppose that X feature-spans \mathcal{X} , Z feature-spans \mathcal{Z} , and z_1, \dots, z_M are feature-independent. Then $f \in I(\mathcal{X})_{\leq d}^\perp$ if and only if there is a vector (c_1, \dots, c_M) in the left range of $K_{\leq d}(X, Z) \cdot K_{\leq d}(Z, Z)^{-1/2}$ such that $f = \sum_{j=1}^M c_j k_{\leq d}(z_j, \mathbf{t})$.*

The statements of this theorem can be turned into statements about $K_{\leq d}(X, X)$ and $K_{\leq d}(Z, Z)$ by taking $\mathcal{X} = \mathcal{Z}$ and $X = Z$. Let us denote the Moore-Penrose pseudoinverse of a matrix A by A^+ . The cross-kernel matrix allows us to reconstruct the kernel matrix $K_{\leq d}(X, X)$ as follows.

Theorem 3.4. *In the above setting we have $K_{\leq d}(X, X) = K_{\leq d}(X, Z) \cdot K_{\leq d}(Z, Z)^+ \cdot K_{\leq d}(Z, X)$ if and only if $\text{fspan}(X) \subseteq \text{fspan}(Z)$. In particular, equality holds if Z feature-spans \mathcal{Z} .*

The equality in Theorem 3.4 is an exact matrix decomposition of the kernel matrix of X which does not require Z to be subsampled from X . This differs markedly from Nyström type methods.

4. The ideal-kernel duality and kernel learning

We demonstrate how kernel-ideal duality and cross-kernel matrices can be employed for common learning tasks. They enable us to improve computational cost, stability, and to obtain novel information about the data manifold. In the setting of the preceding section, we collect some basic observations on the matrix $K := K_{\leq d}(X, Z) \cdot K_{\leq d}(Z, Z)^{-1/2}$ under the assumption that Z feature-spans \mathcal{Z} (or \mathbb{R}^n).

Remark 4.1. (a) Assuming that arithmetic and kernel function evaluation can be done at cost $O(1)$, it takes $O(M^3 + MN)$ steps to compute K and $O(M^2N + M^3)$ steps to obtain a singular value decomposition of K . Thus both tasks require asymptotically linear time in N .

(b) By Theorem 3.4, the equality $K_{\leq d}(X, X) = K \cdot K^\top$ holds *exactly*. Thus a SVD for $K_{\leq d}(X, X)$ can be obtained from a SVD $K = USV^\top$ of K via $K_{\leq d}(X, X) = US^2U^\top$. By (a), the cost of this method is only $O(M^2N + M^3)$, which is linear in N , while the direct algorithm costs $O(N^3)$.

(c) If Z is chosen to be a subsample of X , one recovers common subsampling strategies. However, the feature generating matrix Z can be chosen completely independently from X . Hence the matrix $K_{\leq d}(Z, Z)^{-1/2}$ needs to be precomputed only once for a given kernel degree and data dimension. In particular, it is not necessary to subsample X . For instance, we can choose Z randomly. Or, if desired, the matrix Z can be chosen to contain both a subsample of X and random points.

(d) As mentioned in Theorem 3.3.b, the right singular vectors of K characterize the data manifold \mathcal{X} from which X is sampled. These right singular vectors cannot be obtained from $K_{\leq d}(X, X)$ alone.

The Ideal PCA (IPCA) algorithm

The kernel-ideal duality can be used to both speed up kernel PCA and additionally yield feature functions cutting out the data manifold. By using the matrix $K_{\leq d}(X, Z)$ instead of $K_{\leq d}(X, X)$, we obtain the same feature projections and principal components as in ordinary kernel PCA (both in terms of the X - and the Z -basis), the corresponding singular values, and novel feature projections which cut out the data manifold (which we call “right principal components”). All of this is achieved by Algorithm 1 in linear time and without the necessity of subsampling.

For dimension reduction tasks, the important quantities are evaluations of the principal components. The left principal components agree with the ones from PCA, while the orthogonal of the space generated by the right principal components is used in Algorithm 2 to project onto the data manifold.

The vector u contains exactly the principal feature evaluations of x one would obtain from kernel PCA. The vector v is the IPCA-type right analogue, while v^\perp can be seen as a measure for how closely x lies on the data manifold. Neither v nor v^\perp can be obtained from any algorithm involving only the matrix $K_{\leq d}(X, X)$.

5. Experimental Validation

We validate the main claims of our paper experimentally. We show, in this section: (a) the left principal IPCA features obtained from the cross-kernel are equal to the principal features obtained from kernel PCA [5] and the kernel matrix K_{XX} (5) but faster to obtain, (b) the vanishing

Algorithm 1 IPCA Computes left and right principal vectors of the cross-kernel.

Input: a degree $d \geq 1$, a data matrix $X \in \mathbb{R}^{N \times n}$, a matrix $Z \in \mathbb{R}^{M \times n}$. A threshold $\varepsilon > 0$ or a cut-off m .

Output: m left principal components as columns of a matrix $U \in \mathbb{R}^{N \times m}$ and m right principal components as columns of a matrix $V \in \mathbb{R}^{M \times m}$, as well as the corresponding singular values on the diagonal of a matrix $S \in \mathbb{R}^{m \times m}$.

- 1: Compute the matrices $K_{XZ} = K_{\leq d}(X, Z) \in \mathbb{R}^{N \times M}$ and $K_{ZZ} = K_{\leq d}(Z, Z) \in \mathbb{R}^{M \times M}$.
 - 2: Compute $K = K_{XZ} \cdot K_{ZZ}^{-1/2}$.
 - 3: If centering is desired, replace K by $K - \frac{1}{N} \mathbf{1}K$, where $\mathbf{1} \in \mathbb{R}^{N \times N}$ denotes the all-ones matrix.
 - 4: Compute the truncated singular value decomposition USV^\top of K , with $U \in \mathbb{R}^{N \times m}$, $V \in \mathbb{R}^{M \times m}$, and $S = \text{diag}(\sigma_1, \dots, \sigma_m)$. (If ε is given, then σ_m is the smallest singular value with $\sigma_m \geq \varepsilon$.)
 - 5: Output U, V, S .
-

Algorithm 2 eval-IPCA Computes IPCA features of the data.

Input: a degree $d \geq 1$, a data matrix $X \in \mathbb{R}^{N \times n}$, a matrix $Z \in \mathbb{R}^{M \times n}$, a threshold $\varepsilon > 0$, and a test data point $x_t \in \mathbb{R}^n$.

Output: left principal features $u \in \mathbb{R}^m$ (PCA-like), right principal features $v \in \mathbb{R}^m$ (PCA-like) and certifying features $v^\perp \in \mathbb{R}^m$ (IPCA-like). The entries of u and v are local, whitened coordinates on the data manifold \mathcal{X} (u in X - and v in Z -basis), and the entries of v^\perp are orthogonal coordinates on \mathcal{X} (i.e., they should be close to zero for $x_t \in \mathcal{X}$).

- 1: Apply the IPCA algorithm to obtain matrices U, S, V . The columns of $U \in \mathbb{R}^{N \times m}$ contain m left principal vectors and the columns of $V \in \mathbb{R}^{M \times m}$ contain m right principal vectors.
 - 2: Compute the evaluation vectors $\kappa_X = K(x_t, X)$ and $\kappa_Z = K(x_t, Z)$.
 - 3: Compute (or pass over from IPCA) the matrix $K_{ZZ}^{-1/2} = K_{\leq d}(Z, Z)^{-1/2}$.
 - 4: Compute $P = I - VV^\top$, where I is the identity matrix of size M .
 - 5: Output: $u = \kappa_X \cdot U \cdot S^{-1}$ $v = \kappa_Z \cdot V \cdot S^{-1}$ $v^\perp = \kappa_Z \cdot K_{ZZ}^{-1/2} \cdot P$
-

features obtained from IPCA can be used for manifold learning with kernels, and (c) the IPCA vanishing features outperform kernel PCA features in the USPS classification scenario in terms of compactness, computational cost, and accuracy.

Data sets. Our synthetic data below are: (i) points uniformly from a circle of radius 10, then perturbed by independent 2-dimensional standard, Gaussian random vectors, and (ii) 2 circles on a sphere of radius 5. For each circle, 200 points are sampled uniformly, then 3-dimensional Gaussian noise with variance 0.1 is added. The real world data set used below is (iii) the USPS handwritten digits data set [8]. All experiments below are done with the inhomogenous polynomial kernel of degree 2, with $\theta = 1$.

Kernel PCA vs IPCA. We compared kernel PCA features and IPCA left features on the (ii) 2 circles data. We fixed $M = 12$ as the number of feature spanning points, and computed principal components for $N = 10, 20, \dots, 1000$ data points. The six non-degenerate principal vectors and their principal values were equal to the IPCA left feature vectors and the squares of the corresponding singular values, up to machine precision. Figure 1 shows log-runtimes in comparison. The runtime for IPCA is lower by orders of magnitude.

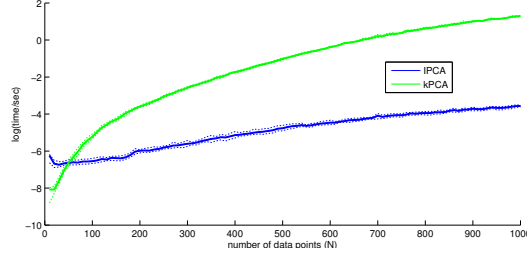


Figure 1: Runtime of kernel PCA and IPCA. Thick line is mean, dotted lines are 0.1 and 0.9 quantiles over 20 repetitions.

Manifold learning with cross-kernel and IPCA. We used IPCA to learn the data manifold from 200 samples in the (i) circle and the (ii) 2 circle example. The manifold is approximated by the points with low evaluation of the normalized IPCA certifying features, the result can be seen in figure 2. The singular value threshold was fixed to $m = 5$ for (i) and $m = 9$ for (ii). The certifying features were evaluated on a fine grid, grid points with evaluation norm near the training data were considered in-manifold, with 0.1-quantile for (i) and 0.01-quantile for (ii).

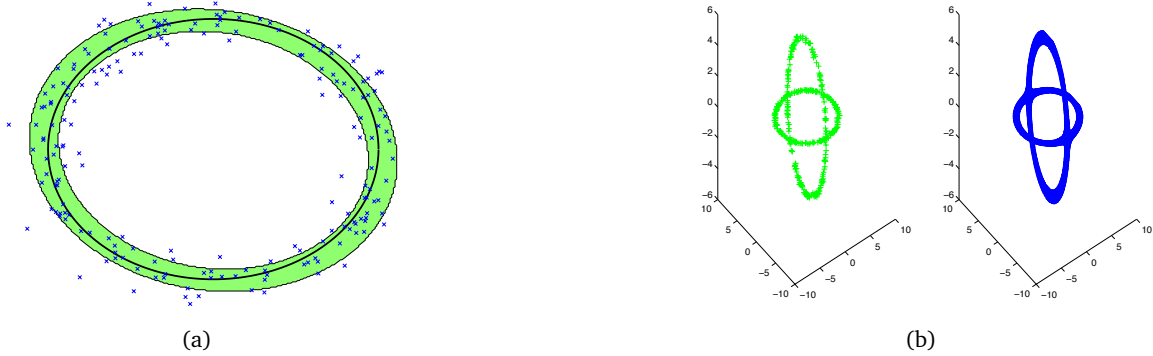


Figure 2: Manifold estimation with IPCA. The crosses are the data points. The estimated manifolds are the light green ring in 2(a) and the two blue rings in 2(b).

Classification with IPCA certifying vectors. We used the IPCA features for classification on the USPS data set. In all experiments, 700 training examples are used, and the rest of the USPS data set is used for evaluation.

Denote the different training classes X_1, \dots, X_t and the pooled data by $X_{\text{train}} := \cup_{i=1}^t X_i$. The natural form of a one-vs-all classifier using IPCA certifying features is as follows: (1) to train, compute $\text{IPCA}(X_i, Z)$ for each of the X_i with the same fixed Z ; to classify a set of points X_{test} , compute the right-evaluation vectors using Algorithm 2 for each class and assign the label corresponding to the minimum norm evaluation. The correctness of this construction follows directly from Theorem 3.3. All the stability and convergence guarantees of IPCA go through without modification.

We compare three choices for the feature generating points Z : (1) choose the feature-generating matrix Z with standard normal Gaussian entries; (2) choose it as a uniform random sub-sample of X_{train} ; (3) choose it as a degenerate sub-sample of X_{train} ; we simulate a high

number of repetitions by sub-sampling from a sub-sample of size $N/4$.

Figure 3(a) summarizes the results. Non-degenerate sub-sampling (2) is best, since it takes advantage of the implicit dimension-reduction of expressing the \mathcal{X}_i in a coordinate system derived $\mathcal{X} \subset \mathbb{R}^n$, which can be of much lower dimension than n . However, when the sampling is *degenerate* as in (3), e.g., with many repetitions, a random feature-generating matrix of the same size as in (1) gives better performance. Eventually, with growing M all choices of Z yield comparable results. As it can not be distinguished a-priori whether the training sample is degenerate or not, random Z can be preferable.

We also compared the IPCA certifying features to the IPCA left and right principal features (PCA-like), by extracting $m = 32$ features, then performing linear classification via LIBLINEAR [3]. The feature-generating matrix Z is chosen randomly with standard normal entries. Figure 3(b)–(c) show that $m = 32$ certifying features yields more accurate and faster results than the same number of principal features. The results reported in [5] for classical kernel PCA are comparable to the PCA-like features, which all are outperformed by the IPCA certifying features.

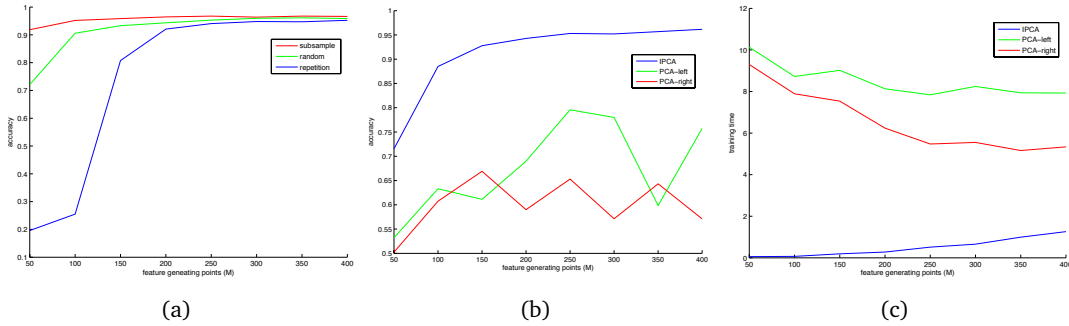


Figure 3: (a) Classification with IPCA certifying features, comparing random Z and subsampled Z from generic and degenerate sampling; (b) performance comparison of the three IPCA features - left and right principal features are closely related to kernel PCA features; (c) speed comparison of IPCA feature based classifiers.

Above comparing different modes of instantiating IPCA, the experiments also show that IPCA is capable of extracting competitive features for considerably less and better scaling computational cost than kernel PCA, and achieves comparable performance on a much smaller number of relevant features.

Acknowledgments LT is supported by the European Research Council under the European Union’s Seventh Framework Programme (FP7/2007-2013) / ERC grant agreement no 247029-SDModels. This research was conducted while MK and LT were guests of FK at Mathematisches Forschungsinstitut Oberwolfach, supported by FK’s Oberwolfach Leibniz Fellowship.

References

- [1] Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. In *Automation and Remote Control*, number 25 in *Automation and Remote Control*, pages 821–837, 1964.

- [2] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.
- [3] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874, 2008.
- [4] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels*. MIT Press, 2002.
- [5] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.
- [6] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, 2004.
- [7] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
- [8] USPS handwritten digits data set *A Database for Handwritten Text Recognition Research*, J. J. Hull, *IEEE PAMI* 16(5) 550-554, 1994.

Appendix

Proof: (of Proposition 2.3)

Let $m = \binom{n+d}{d}$. Under the isomorphism $\mathbb{R}[\mathbf{t}]_{\leq d} \rightarrow \mathbb{R}^m$ given by $\sum_{\alpha} c_{\alpha} \mathbf{t}^{\alpha} \mapsto (c_{\alpha})_{|\alpha| \leq d}$, the polynomial $\mathbf{k}_{\leq d}(x, \mathbf{t})$ maps to $(\gamma_{\alpha}^2 x^{\alpha})_{|\alpha| \leq d}$ for every $x \in \mathbb{R}^n$. Up to scaling the coordinates by γ_{α}^2 , the image of this isomorphism is a Veronese variety. Now the claim follows from the fact that this Veronese variety is irreducible and non-degenerate. \square

Proof: (of Proposition 2.4)

First we show (a). Both vector spaces have the same dimension. The map φ is \mathbb{R} -linear and sends the basis vector \mathbf{t}^{α} to the basis vector $(1/\gamma_{\alpha}) \cdot e_{\alpha}^*$, i.e., to the map defined by $e_{\beta} \mapsto (1/\gamma_{\alpha}) \delta_{\alpha\beta}$. From this the claim follows.

To prove (b), we use the standard scalar product on $\mathcal{F}_{\leq d}$ to identify it with $(\mathcal{F}_{\leq d})^{\vee}$. This (non-canonical) isomorphism identifies the standard basis vector e_{α} with e_{α}^* . Thus the claim follows from (a). \square

Proof: (of Corollary 2.5)

It is easy to check that $\Phi_{\leq d}^* = \text{eval} \circ \varphi^{-1}$. Since both maps eval and φ^{-1} are isomorphisms, the claim follows. \square

Proof: (of Proposition 2.6)

This follows by writing $f = \sum_{\alpha} c_{\alpha} \mathbf{t}^{\alpha}$ with $c_{\alpha} \in \mathbb{R}$ and computing

$$\langle f, \mathbf{k}_{\leq d}(x, \mathbf{t}) \rangle_{\varphi} = \langle \sum_{\alpha} c_{\alpha} \mathbf{t}^{\alpha}, \sum_{\beta} \gamma_{\beta}^2 x^{\beta} \mathbf{t}^{\beta} \rangle_{\varphi} = \sum_{\alpha} c_{\alpha} \gamma_{\alpha}^2 x^{\alpha} \langle \mathbf{t}^{\alpha}, \mathbf{t}^{\alpha} \rangle_{\varphi} = \sum_{\alpha} c_{\alpha} x^{\alpha} = f(x)$$

\square

Proof: (of Theorem 2.8)

Part (a) follows from the definition of $I(\mathcal{X})$ and Prop. 2.6.

To prove (b), we first use (a) in order to map $I(\mathcal{X})_{\leq d}$ isomorphically into $(\mathcal{F}_{\leq d})^{\vee}$. The result is the orthogonal complement of $U = \langle \sum_{|\alpha| \leq d} \gamma_{\alpha} x^{\alpha} e_{\alpha}^{\vee} \mid x \in \mathcal{X} \rangle$. Since we are using the standard scalar products on $\mathcal{F}_{\leq d}$ and its dual, this orthogonal complement equals the annihilator of the preimage of U in $\mathcal{F}_{\leq d}$. That preimage is generated by the elements $\sum_{|\alpha| \leq d} \gamma_{\alpha} x^{\alpha} e_{\alpha} = \Phi_{\leq d}(x)$ with $x \in \mathcal{X}$. In other words, the vector space U^{\perp} is the annihilator of the feature span of \mathcal{X} , as claimed.

Finally, we note that (c) follows from (b) by identifying $\mathcal{F}_{\leq d}$ with its dual via the standard scalar product. \square

Proof: (of Theorem 3.3)

First we prove claim (a). Letting $v_i = \Phi_{\leq d}(x_i) \in \mathbb{R}^m$ and $w_j = \Phi_{\leq d}(z_j) \in \mathbb{R}^m$ for $i = 1, \dots, N$ and $j = 1, \dots, M$, we have to show that the rank of the matrix $G = (\langle v_i, w_j \rangle)_{i,j}$ equals the dimension of $\langle v_1, \dots, v_N \rangle$. By removing columns and rows of G , we may assume that $\{v_1, \dots, v_N\}$ and $\{w_1, \dots, w_M\}$ are linearly independent. Using the hypothesis $\langle v_1, \dots, v_N \rangle \subseteq \langle w_1, \dots, w_M \rangle$, we see that $M \geq N$ and may assume $v_i = w_i$ for $i = 1, \dots, N$. Then the first N columns of G are the Gram matrix of $V = (v_1, \dots, v_N)$ and the claim follows from the fact that the rank of the Gram matrix of V equals the rank of V .

Next we show (b). The implication “ \Leftarrow ” is trivially true. To prove the converse, we note that the hypothesis implies that there are $a_1, \dots, a_N \in \mathbb{R}$ such that $\Phi_{\leq d}(c) - \sum_{i=1}^N a_i \Phi_{\leq d}(x_i)$ is orthogonal to $\text{fspan}(Z)$. Since this vector is contained in $\text{fspan}(Z)$, it is zero. To prove the additional claim, we note that $\Phi_{\leq d}(c) \in \text{fspan}(X)$ implies that all linear forms in the annihilator

of $\text{fspan}(X)$ vanish at c . Via φ^{-1} , these linear forms correspond to the polynomials in $I(\mathcal{X})_{\leq d}$. Since \mathcal{X} is cut out by these polynomials, we get $c \in \mathcal{X}$, as claimed.

For the proof of part (c) we use the isomorphism φ to write the hypothesis $f \in I(\mathcal{Z})_{\leq d}^\perp$ as $f = \sum_{j=1}^M c_j \mathbf{k}_{\leq d}(z_j, \mathbf{t})$ for some $c_j \in \mathbb{R}$. We have $f \in I(\mathcal{X})$ if and only if $\varphi(f)$ is orthogonal to $\langle \varphi(\mathbf{k}_{\leq d}(x_i, \mathbf{t})) \rangle$. Therefore we require that we have

$$\begin{aligned} 0 &= \langle \varphi(f), \varphi(\mathbf{k}_{\leq d}(x_i, \mathbf{t})) \rangle = \langle \sum_{j=1}^M c_j \sum_{\alpha} \gamma_{\alpha} z_j^{\alpha} e_{\alpha}^{\vee}, \sum_{\alpha} \gamma_{\alpha} x_i^{\alpha} e_{\alpha}^{\vee} \rangle \\ &= \sum_{\alpha} \gamma_{\alpha}^2 \left(\sum_{j=1}^M c_j z_j^{\alpha} \right) x_i^{\alpha} \end{aligned}$$

for $i = 1, \dots, N$. On the other hand, the entry in position (i, j) of $K_{\leq d}(X, Z)$ is $k_{\leq d}(x_i, z_j) = \langle \Phi_{\leq d}(x_i), \Phi_{\leq d}(z_j) \rangle = \langle \sum_{\alpha} \gamma_{\alpha} x_i^{\alpha} e_{\alpha}, \sum_{\alpha} \gamma_{\alpha} z_j^{\alpha} e_{\alpha} \rangle = \sum_{\alpha} \gamma_{\alpha}^2 x_i^{\alpha} z_j^{\alpha}$. Hence $(c_1, \dots, c_M)^{\top}$ is in the nullspace of $K_{\leq d}(X, Z)$ if and only if $\sum_{j=1}^M \sum_{\alpha} \gamma_{\alpha}^2 x_i^{\alpha} z_j^{\alpha} c_j = 0$ for $i = 1, \dots, N$. As we have seen, this is equivalent to $f \in I(\mathcal{X})$.

Finally we show (d). Using Theorem 2.8, we see that the hypothesis yields $I(\mathcal{X})^\perp = \langle \mathbf{k}_{\leq d}(x_i, \mathbf{t}) \mid i = 1, \dots, N \rangle$ and $I(\mathcal{Z})^\perp = \langle \mathbf{k}_{\leq d}(z_j, \mathbf{t}) \mid j = 1, \dots, M \rangle$. Therefore there exist numbers $c_j \in \mathbb{R}$ such that we have $f = \sum_{j=1}^M c_j \mathbf{k}_{\leq d}(z_j, \mathbf{t})$, and the question is whether there exists a representation $f = \sum_{i=1}^N a_i \mathbf{k}_{\leq d}(x_i, \mathbf{t})$ with $a_i \in \mathbb{R}$. By the hypothesis, the matrix $K_{\leq d}(Z, Z)$ is invertible, since it is the Gram matrix of the full rank matrix $(\Phi_{\leq d}(z_j))$. So, knowing (c_1, \dots, c_M) is equivalent to knowing

$$(c_1, \dots, c_M) \cdot K_{\leq d}(Z, Z) = \left(\sum_{\alpha} \gamma_{\alpha}^2 \left(\sum_{\ell=1}^M c_{\ell} z_{\ell}^{\alpha} \right) z_j^{\alpha} \right)_j$$

To prove the implication “ \Rightarrow ” we can write the latter tuple as

$$\left(\sum_{\alpha} \gamma_{\alpha}^2 \left(\sum_{i=1}^N a_i x_i^{\alpha} \right) z_j^{\alpha} \right)_j = (a_1, \dots, a_N) \cdot K_{\leq d}(X, Z)$$

and conclude that $c = (a_1, \dots, a_N) \cdot K_{\leq d}(X, Z) \cdot K_{\leq d}(Z, Z)^{-1}$. For the implication “ \Leftarrow ”, we are given the equalities

$$\sum_{\alpha} \gamma_{\alpha}^2 \left(\sum_{\ell=1}^M c_{\ell} z_{\ell}^{\alpha} \right) z_j^{\alpha} = \sum_{\alpha} \gamma_{\alpha}^2 \left(\sum_{i=1}^N a_i x_i^{\alpha} \right) z_j^{\alpha}$$

for $j = 1, \dots, M$. This means that, for $j = 1, \dots, M$, the vectors $\sum_{\ell=1}^M \gamma_{\alpha} c_{\ell} z_{\ell}^{\alpha} e_{\alpha} - \sum_{i=1}^N \gamma_{\alpha} a_i x_i^{\alpha} e_{\alpha}$ and $\sum_{j=1}^M \gamma_{\alpha} z_j^{\alpha} e_{\alpha}$ are orthogonal with respect to the standard scalar product in $\mathcal{F}_{\leq d}$. Thus the first vector is both orthogonal to $\text{fspan}(Z)$ and contained in $\text{fspan}(Z)$, i.e., it is zero. Consequently, we get $\sum_{\ell=1}^M c_{\ell} z_{\ell}^{\alpha} = \sum_{i=1}^N a_i x_i^{\alpha}$, and as we have seen, this yields $f \in I(\mathcal{X})_{\leq d}^\perp$, as claimed. \square

Proof: (of Theorem 3.4)

Let $Y \in \mathbb{R}^{(N+M) \times n}$ be the matrix obtained by row concatenating X and Z . Observe that, by construction, there are projection matrices P and Q such that $K(X, X) = P \cdot K(Y, Y) \cdot P^{\top}$ and $K(X, Z) = P \cdot K(Y, Y) \cdot Q^{\top}$, as well as $K(Z, Z) = Q \cdot K(Y, Y) \cdot Q^{\top}$.

To prove the implication “ \Rightarrow ”, we suppose that $\text{fspan}(X) \not\subseteq \text{fspan}(Z)$. By the rank-nullity theorem, this implies $\text{frk}(Y) > \text{frk}(Z)$. Thus Theorem 3.3.a yields $\text{rank } K_{\leq d}(Z, Z) < \text{rank } K_{\leq d}(Y, Y)$. The claim then follows by applying Lemma 5.1 to $A = K_{\leq d}(Y, Y)$ and $B = Q$ and $C = Q^{\top}$.

For the reverse implication, we first note that $\text{fspan}(X) \subseteq \text{fspan}(Z)$ implies $\text{fspan}(Y) = \text{fspan}(Z)$. Thus we have $\text{frk}(Y) = \text{frk}(Z)$ and Theorem 3.3.a shows $\text{rank } K_{\leq d}(Z, Z) = \text{rank } K_{\leq d}(Y, Y)$. Now the claim follows by applying Lemma 5.1 in the same way as before. \square

The following Lemma provides the key ingredient for the preceding proof.

Lemma 5.1. *Let $A \in \mathbb{R}^{m \times n}$, let $B \in \mathbb{R}^{n \times k}$, and let $C \in \mathbb{R}^{\ell \times m}$. Then the matrix equality $A = AB(CAB)^+CA$ holds if and only if $\text{rank} A = \text{rank} CAB$.*

Proof: First we show the implication “ \Leftarrow ”. Since $\text{rank} CAB = \text{rank} A$, we have $\text{rank} CA = \text{rank} AB = \text{rank} A$, and therefore $\text{colspan}(CA) = \text{colspan}(A)$ as well as $\text{rowspan}(AB) = \text{rowspan}(A)$. This implies $A = C^+CA = ABB^+$. The definition of the pseudoinverse yields $CAB = CAB(CAB)^+CAB$. Multiplying this equality by C^+ and B^+ , we get $A = C^+CABB^+ = C^+CAB(CAB)^+CABB^+ = AB(CAB)^+CA$, as claimed.

To prove the reverse implication, suppose that $\text{rank} A \neq \text{rank} CAB$. Since the rank of a matrix does not increase by matrix multiplication, we must have $\text{rank} A > \text{rank} CAB$. Since the pseudoinverse of CAB satisfies $\text{rank}(CAB)^+ = \text{rank} CAB$, we find $\text{rank} A > \text{rank} CAB = \text{rank}(CAB)^+ \geq \text{rank}(AB(CAB)^+CA)$, in contradiction to the hypothesis. \square